Navigating the Seismic Shift of Post-Moore Computer Systems Design

Anindya Banerjee, Sankar Basu, Erik Brunvand, Pinaki Mazumder, Rance Cleaveland, Gurdip Singh, Margaret Martonosi, and Fernanda Pembleton, *U.S. National Science Foundation, Alexandria, VA, 22314, USA*

In quick succession between 1964 and 1971, our field saw the proposal of Moore's law,[1] the coining of the term "computer architecture,"[2] and the introduction of the first microprocessor.[3] For much of the five decades since then, we have benefitted extraordinarily from both the dynamism of Moore's law transistor scaling and the stable durability of the hardware–software abstractions of computer architecture.

The dynamic duo of Moore's law and computer architecture have allowed massive scaling to occur, and also to be navigated smoothly with relatively little software impact. For example, in the late 1980s and early 1990s, surges in power density occurred as we reached challenging limits in very large scale integration (VLSI) designs based on bipolar transistors; a technology transition from bipolar to complementary metal–oxide–semiconductor (CMOS) occurred with relatively little impact or awareness from the software portion of the computing community.[4]

Over the past 10–15 years however, more fundamental shifts have occurred. For example, Dennard scaling,[5] a companion phenomenon to Moore's law stating that power density could remain stable while transistor sizes shrank, is reaching physical limits. This means that further Moore's law increases in transistor counts are becoming more complex and are only achieved with great effort and at higher power-density costs. Furthermore, as we reach fundamental physical limits in the functioning of small semiconductor transistors, Moore's law itself is being challenged by the increased physical effort and financial expense required to maintain transistor scaling trends.

Hardware designers have responded to these post-Dennard and soon post-Moore trends through increased use of parallelism and heterogeneity. A decade or so ago, homogeneous chip multiprocessors employing multiple copies of general-purpose cores were on the upswing. These on-chip parallel approaches acknowledge that often the best power-performance for a given workload occurs when multiple lightweight compute elements are employed (at a lower clock rate) rather than a monolithic large and high-clock-rate single-processor approach.

A second-wave response to post-Dennard and post-Moore trends has been the growing use of specialized accelerators and heterogeneous parallelism. In tailoring compute elements and functional units to specific useful tasks, heterogeneous parallelism can achieve better power-performance than prior approaches, by tailoring hardware to an application's needs and "pruning away" generality that often incurs a power cost.

WE ARE HERE

After years of increasing parallelism and heterogeneity, we are now navigating what has been termed a Cambrian explosion of diverse and specialized hardware elements.[6] From programmable elements like graphics processing units (GPUs) and radio baseband processors to more specialized elements like single-function accelerators, the hardware–software interface must now accommodate a rich and fast-changing set of computational units, a far cry from the durable long-term operational abstractions reflected in the instruction set architectures (ISAs) of 30 and 40 years ago.

This reliance on specialized elements has allowed computer systems to keep pace with the performance needs of cutting-edge software applications but has not come without a cost. In particular, the way we have transitioned to heterogeneity has left the field without the truly durable abstractions of the past. While ISAs still exist and have their role, they cannot form the primary hardware–software abstraction, given that well over half the chip area on most microprocessors is now devoted to non-ISA units like accelerators. In embracing specialization for the near-term power-performance challenges it mitigates, we are accepting in exchange a dramatic loss of software

Digital Object Identifier 10.1109/MM.2021.3114899
Date of current version 19 November 2021.

portability from one platform to another, and from one chip generation to another.

> *IF HARDWARE IS CHANGING QUICKLY AND IN WAYS THAT ARE FULLY OR MOSTLY EXPOSED TO SOFTWARE, SOFTWARE MUST ALSO ADAPT RAPIDLY. THIS FUNDAMENTAL SHIFT IS ALMOST CERTAIN TO LEAD TO RELIABILITY AND SECURITY CHALLENGES.*

Furthermore, if hardware is changing quickly and in ways that are fully or mostly exposed to software, software must also adapt rapidly. This fundamental shift is almost certain to lead to reliability and security challenges. Comprehensive hardware and software systems checks are hard to perform on a complex and fast-moving target. With the explosion of heterogeneous hardware elements, the end-to-end reliability of fundamental software tools like compilers becomes a challenge, as different parts of a program are compiled to different and fast-changing hardware. Despite impressive strides in compiler correctness research,[7] it is far from straightforward to obtain end-to-end verified code generation for compilers.

POST-MOORE SEISMIC SHIFT

We refer to this sequence—from post-Dennard physics challenges to specialized hardware designs, to software portability/reliability challenges—as the post-Moore seismic shift. Where some very major technology shifts have previously been navigated with little software impact, the post-Moore seismic shift is already deeply affecting the full field of computer and information science and engineering, with more challenges to come.

The seismic shift is on display when one counts up dozens of distinct processor and accelerator types on a cutting-edge chip. The seismic shift is also on display when one notes how application programming interfaces (APIs) (like CUDA, TensorFlow, PyTorch, and others) are beginning to supplant ISAs as important hardware–software abstractions. For example, in NVIDIA GPUs, the ISA is typically not externally exposed to software at all; CUDA is the preferred hardware–software interface. It has already been nearly ten years since commercial processors reached a tipping point in which more area is devoted to specialized accelerators than to general-purpose central processing units (CPUs). As an example, for the line of processors used in Apple phones, it has been nearly 10 years already since the area devoted to accelerators represented over half the chip area.[8]

FUNDING THE FUTURE

As a research agency, the U.S. National Science Foundation (NSF) has the role of fostering foundational discoveries and translational impact. One of the roles of the NSF Computer and Information Science and Engineering (CISE) is to anticipate disruptions like the post-Moore Seismic Shift and stimulate the research that is needed to continue the radical computer systems innovation of the past few decades well beyond it. This is done in concert with other investments in other important topic areas, and generally in CISE topic areas overall. We do so in a multifaceted and cross-layer approach.

CISE Core Programs

CISE Core programs fund research across all CISE topic areas. Several of the funding clusters within NSF CISE Core programs pertain to research in these areas.

› The Foundations of Emerging Technologies (FET) cluster within CISE Core programs aims to enable radical innovations across many computer systems and theory areas, through research at the intersection of computing and biological systems, nanoscale science and engineering, quantum information science, and other nascent, yet promising, fields of research. FET's goal is to foster research on the technologies themselves, and also on their natural connections to other parts of the computer design hardware–software ecosystem. Examples include research on Ising machine hardware platforms using superconducting devices (adiabatic quantum computing), coherent optics, and CMOS circuits. Other examples demonstrate novel neuromorphic architectures comprising 2-D materials to outperform CMOS deep learning machines in certain real-time applications. Nonvolatile semiconductor memories (NVSMs) are spin torque transfer random access memory (RAM), resistive RAM, phase change RAM, and ferroelectric RAM. Besides memories and storage systems, CISE has also funded projects to build experimental carbon nanotube (CNT)-based RISC processors by overcoming the intrinsic stochasticity of CNT devices due to their chirality (metallic/insulation property) properties. NSF also funds particular technology areas in depth, such as NVSMs and research on computational uses of

CNT systems. Finally, our Quantum Leap and Quantum Computing and Information Systems (QCIS) Faculty Fellows programs seek to expand research in quantum information science and engineering.

- The Software Hardware Foundations cluster within CISE Core programs focuses on the compiler, architecture, and design automation implications as these technologies are employed in computational systems. This cluster also supports research in other topics relevant to the seismich shift, including programming languages, formal methods, and software engineering.
- The Computer and Network Systems cluster within NSF CISE Core programs emphasizes a system focus and awareness of the interdependency and blurring of boundaries among computing, storage, and networking. This includes the resources from which these systems are built (computing, memory, storage, communication networks, accelerators, etc.), and the software systems that run on that underlying hardware.

Principles and Practices of Scalable Systems (PPoSS)

The objective of the PPoSS program is to build a community of researchers who will work symbiotically to perform basic research on scalability, performance, and correctness and accuracy of modern applications, systems, and toolchains built on heterogeneous architectures. PPoSS expects coordinated progress at the intersection of multiple disciplines including, but not limited to, computer architecture, high-performance computing, programming languages and compilers, machine programming, security and privacy, systems, and theory and algorithms. Cross-cutting concerns such as scalability and performance, correctness and accuracy, and architectural heterogeneity must be addressed from the outset in all aspects of systems design and implementation and must be tackled with respect to the full hardware and software stack.

Secure and Trustworthy Cyberspace (SaTC)

In today's world, cybersecurity involves hardware, software, networks, data, people, and integration with the physical world. Society's overwhelming reliance on this complex cyberspace, however, has exposed its fragility and vulnerabilities that defy existing cyberdefense measures. Achieving a truly secure cyberspace requires addressing both challenging scientific and engineering problems involving many components of a system, and vulnerabilities that stem from human behaviors and choices. Examining the fundamentals of security and privacy as a multidisciplinary subject can lead to fundamentally new ways to design, build and operate cyber systems, protect existing infrastructure, and motivate and educate individuals about cybersecurity. The complex interplay of hardware and software in post-Moore systems adds further challenges to the topic space SaTC must cover.

Designing Accountable Software Systems (DASS)

The DASS program solicits foundational research aimed towards a deeper understanding and formalization of the bidirectional relationship between computer systems and the complex social and legal contexts within which these systems must be designed and operate. Consider, for example, the California Privacy Rights Act, the EU General Data Protection Regulation, and others. These laws place expectations on the underlying computer systems that implement and uphold them; how do we move towards accountable designs? The DASS program aims to bring researchers in computer and information science and engineering together with researchers in law and social, behavioral, and economic sciences to jointly develop rigorous and reproducible methodologies for understanding the drivers of social goals for software and for designing, implementing, and validating accountable software systems. While DASS is initially focused on software systems, one can envision future scenarios calling for full-stack accountability including the hardware–software interface and underlying hardware as well.

Formal Methods in the Field (FMitF)

The FMitF program aims to bring together researchers in formal methods with researchers in other areas of computer and information science and engineering to jointly develop rigorous and reproducible methodologies for designing and implementing correct-by-construction systems and applications with provable guarantees. FMitF encourages close collaboration between two groups of researchers. The first group consists of researchers in the area of formal methods, which is broadly defined as principled approaches based on mathematics and logic to system modeling, specification, design, analysis, verification, and synthesis. The second group consists of researchers in the "field," which is defined as a subset of areas within computer and information science and engineering that currently do not benefit from having established communities already developing and applying formal methods in their research. Currently, the field comprises the following areas that stand to directly benefit from a grounding in formal methods: computer networks, distributed/operating systems, embedded systems, human-centered computing, and machine learning. FMitF encourages new techniques, grounded in formal methods, to overcome the complexity, reliability, and verification challenges posed by post-Moore systems.

Resilient & Intelligent NextG Systems (RINGS)

The RINGS program seeks to accelerate research in areas that will potentially have a significant impact on emerging Next Generation (NextG) wireless and mobile communication, networking, sensing, and computing systems, with a focus on greatly improving the resiliency of such networked systems. These systems can be thought of as a wildly heterogeneous computing/storage system that spans mobile devices, edge computing resources, and cloud computing resources, all connected with NextG network links. The resiliency of such systems, which subsumes security, adaptability, and autonomy, will be a key driving factor for future NextG systems. This program seeks to go beyond the current research portfolio within the individual directorates by funding collaborative teams that can simultaneously emphasize gains in resiliency (through security, adaptability and/or autonomy) across all layers of the networking protocol and computation stacks as well as in throughput, latency, and connection density.

Of particular note in the programs above is the degree to which cross-layer interactions are emphasized either in a general way (PPOSS) or in application-focused ways (RINGS, FMitF, Quantum).

WHAT IS NEXT?

Several of the programs above are general and far-reaching, intending to navigate years of change that lie ahead. In addition to long-term core investments, NSF will continue to envision future programs and invest in them as resources permit and science/technology opportunities dictate.

As a community, it will take more than research funding programs to navigate what lies ahead. In particular, in addition to funding, we see a need for an all-of-CISE push on the following elements:

- Research communities: The full CISE research community must develop plans and activities to cross the Special Interest Group (SIG) boundary lines and create a culture in which more interdisciplinary community meetings and publication methods become the norm. Programs like PPoSS, RINGS, FMitF, and DASS seek to fund collaborations across traditional boundaries; their success hinges on vibrant opportunities for such research communities to form and advance. Likewise, conferences and publication processes should be open and broaden to foster such work.
- Prototyping and open-sourcing: Impactful work on these full-stack and interdisciplinary topics will require ecosystems that form so that research advances can build off each other in a modular way. This calls for additional community attention on prototyping and open-sourcing. Prototyping includes fostering the ability for researchers in our fields to gain access to integrated circuit (IC) fabrication capabilities so that novel architecture and circuits ideas can be tested in real chips. Open-source accelerators and processors can be the building blocks for evaluating and prototyping research on even more complex designs. Well-documented open-source hardware can also serve as the target for open-source software systems that compile, optimize and verify for these full-stack efforts. Success at this may require reassessing aspects of how open-sourcing impacts are accounted for in publication review processes and promotion and tenure (P&T) processes.
- IC fabrication: Full-stack and interdisciplinary research, including advanced accelerator and application-specific architecture studies also may rely upon special-purpose ICs. To maximize the impact of these studies, researchers should have access to IC fabrication facilities (both actual fabrication and also high-fidelity simulation models). Some research will require the most advanced process nodes, whereas for other ideas more reliable and sustainable process nodes may be sufficient, especially where more exotic technologies are

> *WE AT NSF WELCOME PROPOSALS FROM THE U.S. RESEARCH COMMUNITY TO ADVANCE ON BOTH THE TECHNICAL CHALLENGES OUTLINED HERE AND THE ASPECTS OF COMMUNITY-BUILDING AND NEW PUBLICATION METHODOLOGIES THAT WILL BEST POSITION THE COMMUNITY TO TACKLE THEM.*

proposed to be combined with traditional silicon substrates. The CISE community, including industry partners in IC fabrication and design tools, should seek partnerships and opportunities to offer fabrication opportunities to researchers, academics, and perhaps most importantly, students.

- Translation from research to production systems: Finally, attention to translational impact will be key to meeting the post-Moore imperative to create resilient and high-performance, and secure computer systems from the underlying hardware that is complex and heterogeneous. Testing, verification, and prototyping at-scale will be key to determining promising future pathways for post-Moore computer systems design. We as a community need to better support prototyping and translational impact, in all aspects of how our community interacts and operates (the role of conference planners, P&T committees, etc.)

SUMMARY

We at NSF welcome proposals from the U.S. research community to advance on both the technical challenges outlined here and the aspects of community-building and new publication methodologies that will best position the community to tackle them. We likewise look forward to a range of future community partnerships that can help future computing systems in the wake of the post-Moore Seismic Shift to achieve the heightened levels of reliability, security, performance, and portability that society demands.

REFERENCES

1. G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, Apr. 19, 1964. [Online]. Available: https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf
2. G. Amdahl, G. Blaauw and F. Brooks, "Architecture of the IBM System/360," *IBM J. Res. Develop.*, vol. 8, no. 2, pp. 87–101, Apr. 1964.
3. Intel, "The Story of the Intel 4004," [Online]. Available: https://www.intel.com/content/www/us/en/history/museum-story-of-intel-4004.html
4. S. Kaxiras and M. Martonosi, "Computer architecture techniques for power efficiency," in *Synthesis Lectures on Computer Architecture*. San Rafael, CA, USA: Morgan Claypool, 2008.
5. R. Dennard, F. H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, no. 5, pp. 256–268, Oct. 1974.
6. J. L. Hennessy and D. A. Patterson, "A new golden age for computer architecture," *Commun. ACM*, vol. 62, no. 2, pp. 48–60, 2019.
7. X. Leroy, "Formal Verification of a Realistic Compiler," *Commun. ACM*, vol. 52, no. 7, pp. 107–115, 2009.
8. Y. S. Shao, B. Reagen, G. Wei and D. Brooks, "The aladdin approach to accelerator design and modeling," *IEEE Micro*, vol. 35, no. 3, pp. 58–70, May/Jun. 2015. doi: 10.1109/MM.2015.50.

ANINDYA BANERJEE is a Program Director with the Computing and Communications Foundations (CCF) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation, Alexandria, VA, USA. Contact him at ABANERJE@nsf.gov.

SANKAR BASU is a Program Director with the Computing and Communications Foundations (CCF) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation, Alexandria, VA, USA. Contact him at sabasu@nsf.gov.

ERIK BRUNVAND is a Program Director with the Computer and Network Systems (CNS) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation (NSF), Alexandria, VA, USA. He is serving his term at NSF as a rotator on leave from the University of Utah where he has been on the faculty since 1990. Contact him at ebrunvan@nsf.gov.

PINAKI MAZUMDER is a Program Director with the Computing and Communications Foundations (CCF) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation (NSF), Alexandria, VA, USA. He has been a Professor at the University of Michigan since 1987 and joined NSF in 2020 as a rotator to manage the Foundations of Emerging Technologies (FET) program. Contact him at pmazumde@nsf.gov.

RANCE CLEAVELAND serves as Director of the Computing and Communications Foundations (CCF) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation, Alexandria, VA, USA. He is also Professor of computer science with the University of Maryland at College Park, College Park, MD, USA. Contact him at wrcleave@nsf.gov.

GURDIP SINGH is the Division Director for Computer and Network Systems (CNS) Division, Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation, Alexandria, VA, USA. He is on leave from Syracuse University where he was the Associate Dean for Research and Graduate Programs. Contact him at gsingh@nsf.gov.

MARGARET MARTONOSI leads the Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation (NSF), Alexandria, VA, USA. She is serving her term at NSF as a rotator on leave from Princeton University where she has been on the faculty since 1994. Contact her at mmartono@nsf.gov.

FERNANDA PEMBLETON is Communications Specialist for the Computer and Information Science and Engineering (CISE) directorate, U.S. National Science Foundation, Alexandria, VA, USA. Contact her at mpemblet@nsf.gov.